NBDL-80R003

DEVELOPMENT OF PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL RESEARCH (PETER): COMPLEX COUNTING TEST

Robert S. Kennedy and Alvah C. Bittner, Jr.

October 1980

NAVAL BIODYNAMICS LABORATORY
New Orleans, Louisiana

Approved for public release. Distribution unlimited.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NBDL 80 R003	2. GOVT ACCESSION NO.: AD
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): DEVELOPMENT OF PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL RESEARCH (PETER): COMPLEX COUNTING TEST	5. TYPE OF REPORT & PERIOD COVERED: RESEARCH REPORT
	6. PERFORMING ORG. REPORT NUMBER: NBDL 80 R003
7. AUTHOR(s): ROBERT S. KENNEDY ALVAH C. BITTNER, JR.	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: NAVAL BIODYNAMICS LABORATORY BOX 29407 NEW ORLEANS, LA 70189	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: PROJECT# F58524 TASK AREA ZF5852406 WORK UNIT MF58.524-002-5027
11. CONTROLLING OFFICE NAME AND ADDRESS: NAVAL MEDICAL RESEARCH & DEVELOPMENT COMMAND BETHESDA, MD 20014	12. REPORT DATE: FEB 1980
	13. NUMBER OF PAGES: 3
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

INFORMATION PROCESSING, PERFORMANCE TESTING, COGNITION, PETER, TEST BATTERY, COMPLEX COUNTING

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This study is the first in a program to develop a battery of Performance Evaluation Tests for Environmental Research (PETER). Nineteen volunteer subjects were tested daily for 3 weeks on a complex task requiring the operator to keep simultaneous track of several things with changing states. Average daily performances are reported as well as reliabilities of three main types: (1) internal consistency of the test; (2) sensitivity-the ability to differentiate subjects, and (3) stability-consistency of measurement over repeated sessions. The results showed that, on this task, learning was accomplished quickly, and

performance stayed level for 3 weeks. The cross-trial reliability for this test was found relatively stable after 3 d of practice, with a decline of only r=.94 to r=.79 over 11 d. This task is further noted as having several characteristics which make it particularly suitable for use in environmental research. It is concluded that the complex counting test can be recommended for use in environmental and other time-course research.

NBDL-80R003

DEVELOPMENT OF PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL RESEARCH (PETER): COMPLEX COUNTING TEST

Robert S. Kennedy

and

Alvah C. Bittner, Jr.

October 1980

Bureau of Medicine and Surgery
Work Unit MF58.524-002-5027

Approved by

Channing L. Ewing, M. D.
Scientific Director

Released by

Captain J. E. Wenger MC USN
Commanding Officer

Naval Biodynamics Laboratory
Box 29407
New Orleans, LA 70189

SUMMARY PAGE

THE PROBLEM

Advances in engineering technology generate new environments for the human operator. A most important question is whether these environments will disrupt performance. The problem is complicated by the fact that operators often remain in their environments for extended periods of time. Even though performance test batteries have been previously developed, none have been appropriately standardized for extensive repetitions. To this end, an experimental program has been initiated for developing a Performance Evaluation Test for Environmental Research at the Naval Biodynamics Laboratory.

FINDINGS

This study is the first in a program to develop a battery of Performance Evaluation Tests for Environmental Research (PETER). Nineteen volunteer subjects were tested daily for 3 weeks on a complex task requiring the operator to keep simultaneous track of several things with changing states. Average daily performances are reported as well as reliabilities of three main types: (1) internal consistency of the test; (2) sensitivity-the ability to differentiate subjects, and (3) stability-consistency of measurement over repeated sessions. The results showed that, on this task, learning was accomplished quickly, and performance stayed level for 3 weeks. The cross-trial reliability for this test was found relatively stable after 3 d of practice, with a decline of only $r=.94$ to $r=.79$ over 11 d. This task is further noted as having several characteristics which make it particularly suitable for use in environmental research.

RECOMMENDATIONS

It is concluded that the complex counting test can be recommended for use in environmental and other time-course research.

ACKNOWLEDGEMENTS

The authors are indebted to Ms. M. Harbeson and Ms. M. Krause for collecting the data.

Trade names of materials or products of commercial or nongovernment organizations are cited only where essential to precision in describing research procedures or evaluation of results. Their use does not constitute official endorsement or approval of the use of such commercial hardware or software.

Development of Performance Evaluation Tests for Environmental Research (PETER): Complex Counting Test

ROBERT S. KENNEDY and ALVAH C. BITTNER, JR.

Naval Aerospace Medical Research Laboratory Detachment, New Orleans, Louisiana 70189

KENNEDY, R. S., and A. C. BITTNER, JR. *Development of performance evaluation tests for environmental research (PETER): Complex counting task.* Aviat. Space Environ. Med. 51(2): 142-144, 1980.

This study is the first in a program to develop a battery of Performance Evaluation Tests for Environmental Research (PETER). Nineteen volunteer subjects were tested daily for 3 weeks on a complex task requiring the operator to keep simultaneous track of several things with changing states. Average daily performances are reported as well as reliabilities of three main types: 1) internal consistency of the test; 2) sensitivity—the ability to differentiate subjects, and 3) stability—consistency of measurement over repeated sessions. The results showed that, on this task, learning was accomplished quickly, and performance stayed level for 3 weeks. The cross-trial reliability for this test was found relatively stable after 3 d of practice, with a decline of only r=.94 to r=.79 over 11 d. This task is further noted as having several characteristics which make it particularly suitable for use in environmental research. It is concluded that the complex counting test can be recommended for use in environmental and other time-course research.

ADVANCES IN ENGINEERING technology generate new environments for the human operator. A most important question is whether these environments will disrupt performance. The problem is complicated by the fact that operators often remain in their environments for extended periods of time. Even though performance test batteries have been previously developed, none have been appropriately standardized for extensive repetitions. To this end, an experimental program has been initiated for developing a Performance Evaluation Test for Environmental Research (PETER) at the Naval Aerospace Medical Research Laboratory Detachment (5). A complex counting task containing many of the task characteristics identified as important for environmental time-course studies (5) was selected as the first in a series to be studied for possible inclusion in PETER. This task was reviewed by Kennedy and Bruns (7) and was expected to have minimal changes in mean performance over trials. However, the effects of practice on the standard deviation and on reliability were unknown. The present investigation was directed at characterizing the effects of extensive practice on average performance, variance, and reliability.

MATERIALS AND METHODS

Subjects: The subjects were a group of 19 Navy enlisted men, ages 19-24, who had served as volunteer research subjects since induction into the Navy (approximately 18 months). All volunteer subjects were recruited, evaluated, and employed in accordance with procedures specified in Secretary of the Navy Instruction 3900.39 and Bureau of Medicine and Surgery Instruction 3900.6. These instructions are based upon voluntary informed consent, and meet provisions of prevailing national and international guidelines. Although representative of the enlisted Navy population in size and intelligence, subjects were mentally and physically screened to be qualified for hazardous duty environmental research. Subjects were under continuous medical supervision. They were physically fit and well motivated to perform. For a detailed description of the selection procedure, see Thomas, Majweski, Ewing, and Gilbert (9).

Apparatus and Procedure: Three tones were recorded and played back on a Teac Model A-4010SU reel-to-reel tape recorder with a Realistic SA 101 Solid State Amplifier. The tones (100 Hz, 900 Hz and 1800 Hz) which regularly occurred 5, 6, and 8 times/min, respectively, appeared random to the subjects. The recording was produced using three cams attached to a constant-speed, 1 r.p.m. motor and is described in detail elsewhere (4). The auditory signals were heard by the subjects through Realistic NOVA 10 Stereo Headphones at a comfortable listening level (ca. 60 dB). Subjects, seated at desks, held switches with three buttons marked Low (L), Middle (M), and High (H) to correspond to 100 Hz, 900 Hz and 1800 Hz tones, respectively. Subjects' responses for the "L" and "M" buttons were recorded on instrument chart paper on a Techni-Rite Electronics Recorder set at a tape speed of 1 mm/s. The subjects were tested for 15 consecutive weekdays in

The opinions are those of the authors and do not necessarily reflect those of the Department of the Navy.

From Naval Medical Research and Development Command Work Unit MF58.524-002-5027.

groups of four. In the initial experimental session, they heard taped instructions and were then given a 5-min practice period and required to count the occurrence of the low tone only, pushing the response key marked "L" after every fourth low tone. This was performed in order to be certain that the subjects understood the task before continuing. For the 15-min experimental session, the subjects were then instructed to continue to monitor the low tone but also to push the "M" button after they counted every fourth middle tone; and continue to ignore the high tone. This is referred to as Two-Channel Monitoring. On subsequent days, the subjects were given a 1-min warm-up on both tones, followed by a 15-min experimental period. Percent correct scores were obtained for each subject for each of three 5-min segments (4).

RESULTS

The data from this study were analyzed in two phases. During the first phase, an analysis of variance (ANOVA) of mean percent correct performance was conducted, and studies of changes of the percent correct mean and standard deviations were made by graphical analysis. These analyses were performed in order to examine the simple effects of practice on the mean and standard deviation. During the second phase, the effect of practice on the reliability of the test was made by graphical analyses. The results from the two phases of analyses are described below.

Effects of Practice on Mean and Standard Deviation: The effects of practice on mean percent correct are depicted graphically in Fig. 1. A subjects-by-days ANOVA with one observation per call showed a nonsignificant day (practice) effect, ($p=0.5$). Fig. 1, which shows mean percent correct over trials, gives visual confirmation of the ANOVA result that practice had no significant effect on mean percent correct performance. It should be noted that there was also a very highly significant subjects effect ($p<0.005$). This factor, coupled with the inter-day reliabilities discussed later, suggests a very useful test over 15 day's practice, with over 77% of the explained variation accounted for by this effect. However, reliability estimates from the subject effects are only valid if the standard deviations of performance over trials are constant and correlations between different trials are constant (10), conditions not usually met (1,3,11,12). The effects of practice on the standard deviation of subjects' performance are shown in Fig. 1. It may be seen that the standard deviation about mean subject performance appears level over sessions. This observation is statistically confirmed by the nonsignificant ratio of the squares of the largest and smallest squared standard deviation (F-MAX = .71, $p<0.10$). The finding of nonsignificant changes in standard deviation parallels that for mean performance and indicates no practice effects over the extent of the experiment.

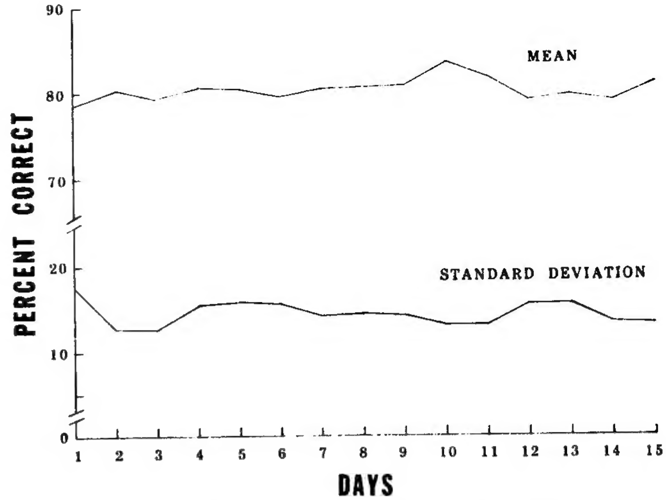

Fig. 1. Mean percent correct and standard deviations over 15 d on a complex counting test (N = 19).

Effects of Practice on Reliability: Table I gives the correlations between all pairs of days across 15 d of practice. Examining this table, it may be noted that reliabilities tend to decline in magnitude as one progresses to the right of the superdiagonal correlations (r_{12}, r_{23}, r_{24}, etc.), in any row. This falloff is illustrated in Fig. 2, where slight downward trends are observed for the correlation between selected base days (1, 2, 4, 9, and 13) and those that follow. This figure is most meaningful when considering the reliability of the test with differing amounts of practice before data collection formally begins (1). In particular, Fig. 2, indicates that base reliabilities for Day 1 tend to be less than those for Base Day 2, and those for Base Day 2 are less than

TABLE I. CORRELATIONS OF PERFORMANCE OVER 15 d OF PRACTICE.

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.70	.61	.65	.60	.64	.78	.55	.76	.78	.60	.53	.50	.51	.59
2		.88	.83	.86	.70	.81	.76	.82	.80	.78	.69	.76	.66	.61
3			.81	.89	.71	.79	.86	.87	.75	.79	.75	.82	.74	.58
4				.94	.88	.87	.87	.85	.86	.85	.86	.80	.79	.79
5					.88	.86	.92	.84	.84	.84	.89	.82	.80	.74
6						.83	.76	.78	.79	.76	.92	.69	.74	.80
7							.85	.88	.87	.78	.79	.78	.76	.69
8								.79	.79	.76	.83	.84	.83	.66
9									.86	.82	.81	.78	.73	.71
10										.71	.79	.65	.69	.78
11											.72	.81	.58	.63
12												.77	.85	.78
13													.83	.63
14														.80
15														

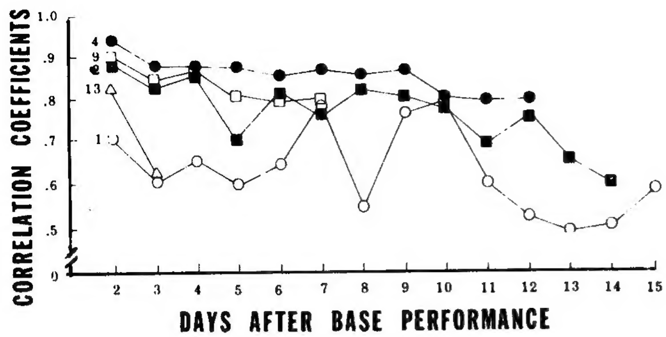

Fig. 2. Relationships between correlations for selected base days (1, 2, 4, 9, and 13) over 15 d on a complex counting test (N = 19).

for Base Day 4; however, Base Day 4 reliabilities are greater than those of Base Day 9, and those for Base Day 9 are greater than for Base Day 13. These findings indicate that giving increasing amounts of practice before beginning formal data collections will result in increased overall reliability, but only up to some point. After reaching this level of performance, a loss occurs due to some aspect of this task. Further examination of this figure shows that, with the possible exception of Base Day 13, more or less parallel trends in correlation decline are seen. This suggests that, regardless of the amount of pretraining on the task, the rate of loss of reliability will be the same. Additional statistical study is needed, however, to delineate the functional relationship between base day, trials and resultant reliability.

DISCUSSION

No effects of practice on the mean and standard deviation of performance were observed in the present study. Although cross trial drops in reliability were seen to occur, they were generally less than what have been found with other tasks investigated by a similar paradigm (6). For example, precipitous drops in reliability from about r=.90 to r=.00 with separation of four trials were shown for a time estimation task. Overall, the complex counting test possesses mean, standard deviation, and relative differential stabilities. It should be noted that modest amounts of practice appear desirable before using the present task in long-term investigations. This is because the reliability trace was seen to maximize after only 3 d of 15-min practice sessions where the subsequent cross trial reliability fell from about r=.94 to r=.79 over 11 d. It is noteworthy that these findings only came to light when the day-to-day reliabilities were graphically followed for purposes of observing trends. If only the average mean performance or composite reliability were used, as suggested by many authors (2), the above relationship with practice would not have been discovered. Because the reliabilities are high and relatively stable over following days, it is felt that this task is suitable for PETER provided that there be three or four practice sessions, and the number of environmental exposures be limited to 6 or 7 d.

In concluding, it is pertinent to note that the complex counting task has been found sensitive to environmental effects in other studies (8) and that it is portable, inexpensive, and easily administered (7). Considering these factors, coupled with the stabilities found above, the complex counting test can be recommended for environmental research and time-course investigations.

REFERENCES

1. Alvares, K. M., and C. L. Hulin. 1972. Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analyses. *Hum. Factors* 14:295-308.
2. Jensen, A. R. 1965. Scoring the Stroop Test. *ACTA Psychologica*. 24:398-408.
3. Jones, M. B. 1966. Individual differences. *In:* E. A. Bilodeau (Ed.), Acquisition of Skills. New York: Academic Press, 109-146.
4. Kennedy, R. A. 1971. A comparison of performance on visual and auditory monitoring tasks. *Hum. Factors* 13:93-97.
5. Kennedy, R. S., and A. C. Bittner, Jr. 1977. The development of a Navy Performance Evaluation Test for Environmental Research (PETER). *In:* Productivity Enhancement: Personnel Performance Assessment in Navy Systems. San Diego, CA: Naval Personnel Research and Development Center. AD#A056047.
6. Kennedy, R. S., and A. C. Bittner, Jr. 1978. Progress in the analysis of a Performance Evaluation Test for Environmental Research (PETER). *Proceedings of the 22nd Annual Meeting of the Human Factors Society*, Detroit, MI. AD#A060676.
7. Kennedy, R. S., and F. A. Bruns. 1975. Some practical considerations for performance testing in exotic environments. *In:* B. O. Hartman (Ed.), Higher Mental Functioning in Operational Environments. Brooks Air Force Base, TX. AGARD Conference Profeedings No. 181 C-4:1-6.
8. Kennedy, R. S., W. F. Moroney, R. M. Bale, H. G. Gregoire, and D. G. Smith. 1972. Motion sickness symptomatology and performance decrements occasioned by hurricane penetrations in C-121, C-130, and P-3 Navy aircraft. *Aerospace Med.* 43:1235-1239.
9. Thomas, D. J., P. L. Majewski, C. L. Ewing, and N. S. Gilbert. 1978. Medical qualification procedures for hazardous-duty aeromedical research. London: AGARD Conference Proceedings No. 231 A3:1-13.
10. Winer, B. J. 1962. Statistical Principles in Environmental Design. New York: McGraw Hill.
11. Woodrow, H. 1938. The effect of practice on groups of different initial ability. *J. Educ. Psychol.* 29:268-278.
12. Woodrow, H. 1938. The effect of practice on test intercorrelations. *J. Educ. Psychol.* 29:561-562.